清华大学优秀博士学位论文丛书

复杂时间序列的统计推断理论及预测方法

李杰（Li Jie）著

Statistical Inference and Forecasting Methods for Complex Time Series

清華大學出版社
北京

内 容 简 介

时间序列模型广泛应用于计量经济学、金融学、生物统计学、工业计量学等领域。本书主要研究了复杂时间序列的理论性质和实际应用，包括对时间序列的分布函数、函数型时间序列，以及局部平稳时间序列多步向前预测区间的统计推断。

本书可作为统计学、数据科学等相关专业本科生或研究生的选修课教材，也可作为统计学科研人员、企业管理人员和国家行政机关工作人员学习预测方法的参考用书。

图书在版编目（CIP）数据

复杂时间序列的统计推断理论及预测方法/李杰著.—北京：清华大学出版社，2023.12
（清华大学优秀博士学位论文丛书）
ISBN 978-7-302-65001-0

Ⅰ. ①复… Ⅱ. ①李… Ⅲ. ①统计数据-数据处理-时间序列分析 Ⅳ. ①O212

中国国家版本馆 CIP 数据核字（2023）第 230806 号

责任编辑：戚 亚
封面设计：傅瑞学
责任校对：赵丽敏
责任印制：沈 露

出版发行：清华大学出版社
网 址：https://www.tup.com.cn, https://www.wqxuetang.com
地 址：北京清华大学学研大厦 A 座 邮 编：100084
社 总 机：010-83470000 邮 购：010-62786544
投稿与读者服务：010-62776969, c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015, zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市东方印刷有限公司
经 销：全国新华书店
开 本：155mm×235mm **印 张：**6.75 **插 页：**4 **字 数：**123 千字
版 次：2023 年 12 月第 1 版 **印 次：**2023 年 12 月第 1 次印刷
定 价：69.00 元

产品编号：101860-01

一流博士生教育
体现一流大学人才培养的高度（代丛书序）[①]

人才培养是大学的根本任务。只有培养出一流人才的高校，才能够成为世界一流大学。本科教育是培养一流人才最重要的基础，是一流大学的底色，体现了学校的传统和特色。博士生教育是学历教育的最高层次，体现出一所大学人才培养的高度，代表着一个国家的人才培养水平。清华大学正在全面推进综合改革，深化教育教学改革，探索建立完善的博士生选拔培养机制，不断提升博士生培养质量。

学术精神的培养是博士生教育的根本

学术精神是大学精神的重要组成部分，是学者与学术群体在学术活动中坚守的价值准则。大学对学术精神的追求，反映了一所大学对学术的重视、对真理的热爱和对功利性目标的摒弃。博士生教育要培养有志于追求学术的人，其根本在于学术精神的培养。

无论古今中外，博士这一称号都和学问、学术紧密联系在一起，和知识探索密切相关。我国的博士一词起源于 2000 多年前的战国时期，是一种学官名。博士任职者负责保管文献档案、编撰著述，须知识渊博并负有传授学问的职责。东汉学者应劭在《汉官仪》中写道："博者，通博古今；士者，辩于然否。"后来，人们逐渐把精通某种职业的专门人才称为博士。博士作为一种学位，最早产生于 12 世纪，最初它是加入教师行会的一种资格证书。19 世纪初，德国柏林大学成立，其哲学院取代了以往神学院在大学中的地位，在大学发展的历史上首次产生了由哲学院授予的哲学博士学位，并赋予了哲学博士深层次的教育内涵，即推崇学术自由、创造新知识。哲学博士的设立标志着现代博士生教育的开端，博士则被定义为

① 本文首发于《光明日报》，2017 年 12 月 5 日。

独立从事学术研究、具备创造新知识能力的人，是学术精神的传承者和光大者。

博士生学习期间是培养学术精神最重要的阶段。博士生需要接受严谨的学术训练，开展深入的学术研究，并通过发表学术论文、参与学术活动及博士论文答辩等环节，证明自身的学术能力。更重要的是，博士生要培养学术志趣，把对学术的热爱融入生命之中，把捍卫真理作为毕生的追求。博士生更要学会如何面对干扰和诱惑，远离功利，保持安静、从容的心态。学术精神，特别是其中所蕴含的科学理性精神、学术奉献精神，不仅对博士生未来的学术事业至关重要，对博士生一生的发展都大有裨益。

独创性和批判性思维是博士生最重要的素质

博士生需要具备很多素质，包括逻辑推理、言语表达、沟通协作等，但是最重要的素质是独创性和批判性思维。

学术重视传承，但更看重突破和创新。博士生作为学术事业的后备力量，要立志于追求独创性。独创意味着独立和创造，没有独立精神，往往很难产生创造性的成果。1929 年 6 月 3 日，在清华大学国学院导师王国维逝世二周年之际，国学院师生为纪念这位杰出的学者，募款修造“海宁王静安先生纪念碑”，同为国学院导师的陈寅恪先生撰写了碑铭，其中写道：“先生之著述，或有时而不章；先生之学说，或有时而可商；惟此独立之精神，自由之思想，历千万祀，与天壤而同久，共三光而永光。”这是对于一位学者的极高评价。中国著名的史学家、文学家司马迁所讲的“究天人之际，通古今之变，成一家之言”也是强调要在古今贯通中形成自己独立的见解，并努力达到新的高度。博士生应该以“独立之精神、自由之思想”来要求自己，不断创造新的学术成果。

诺贝尔物理学奖获得者杨振宁先生曾在 20 世纪 80 年代初对到访纽约州立大学石溪分校的 90 多名中国学生、学者提出：“独创性是科学工作者最重要的素质。”杨先生主张做研究的人一定要有独创的精神、独到的见解和独立研究的能力。在科技如此发达的今天，学术上的独创性变得越来越难，也愈加珍贵和重要。博士生要树立敢为天下先的志向，在独创性上下功夫，勇于挑战最前沿的科学问题。

批判性思维是一种遵循逻辑规则、不断质疑和反省的思维方式，具有批判性思维的人勇于挑战自己，敢于挑战权威。批判性思维的缺乏往往被认为是中国学生特有的弱项，也是我们在博士生培养方面存在的一

个普遍问题。2001 年，美国卡内基基金会开展了一项“卡内基博士生教育创新计划”，针对博士生教育进行调研，并发布了研究报告。该报告指出：在美国和欧洲，培养学生保持批判而质疑的眼光看待自己、同行和导师的观点同样非常不容易，批判性思维的培养必须成为博士生培养项目的组成部分。

对于博士生而言，批判性思维的养成要从如何面对权威开始。为了鼓励学生质疑学术权威、挑战现有学术范式，培养学生的挑战精神和创新能力，清华大学在 2013 年发起“巅峰对话”，由学生自主邀请各学科领域具有国际影响力的学术大师与清华学生同台对话。该活动迄今已经举办了 21 期，先后邀请 17 位诺贝尔奖、3 位图灵奖、1 位菲尔兹奖获得者参与对话。诺贝尔化学奖得主巴里·夏普莱斯（Barry Sharpless）在 2013 年 11 月来清华参加“巅峰对话”时，对于清华学生的质疑精神印象深刻。他在接受媒体采访时谈道：“清华的学生无所畏惧，请原谅我的措辞，但他们真的很有胆量。”这是我听到的对清华学生的最高评价，博士生就应该具备这样的勇气和能力。培养批判性思维更难的一层是要有勇气不断否定自己，有一种不断超越自己的精神。爱因斯坦说：“在真理的认识方面，任何以权威自居的人，必将在上帝的嬉笑中垮台。”这句名言应该成为每一位从事学术研究的博士生的箴言。

提高博士生培养质量有赖于构建全方位的博士生教育体系

一流的博士生教育要有一流的教育理念，需要构建全方位的教育体系，把教育理念落实到博士生培养的各个环节中。

在博士生选拔方面，不能简单按考分录取，而是要侧重评价学术志趣和创新潜力。知识结构固然重要，但学术志趣和创新潜力更关键，考分不能完全反映学生的学术潜质。清华大学在经过多年试点探索的基础上，于 2016 年开始全面实行博士生招生“申请–审核”制，从原来的按照考试分数招收博士生，转变为按科研创新能力、专业学术潜质招收，并给予院系、学科、导师更大的自主权。《清华大学“申请–审核”制实施办法》明晰了导师和院系在考核、遴选和推荐上的权力和职责，同时确定了规范的流程及监管要求。

在博士生指导教师资格确认方面，不能论资排辈，要更看重教师的学术活力及研究工作的前沿性。博士生教育质量的提升关键在于教师，要让更多、更优秀的教师参与到博士生教育中来。清华大学从 2009 年开始探

索将博士生导师评定权下放到各学位评定分委员会，允许评聘一部分优秀副教授担任博士生导师。近年来，学校在推进教师人事制度改革过程中，明确教研系列助理教授可以独立指导博士生，让富有创造活力的青年教师指导优秀的青年学生，师生相互促进、共同成长。

在促进博士生交流方面，要努力突破学科领域的界限，注重搭建跨学科的平台。跨学科交流是激发博士生学术创造力的重要途径，博士生要努力提升在交叉学科领域开展科研工作的能力。清华大学于 2014 年创办了“微沙龙”平台，同学们可以通过微信平台随时发布学术话题，寻觅学术伙伴。3 年来，博士生参与和发起“微沙龙”12 000 多场，参与博士生达 38 000 多人次。“微沙龙”促进了不同学科学生之间的思想碰撞，激发了同学们的学术志趣。清华于 2002 年创办了博士生论坛，论坛由同学自己组织，师生共同参与。博士生论坛持续举办了 500 期，开展了 18 000 多场学术报告，切实起到了师生互动、教学相长、学科交融、促进交流的作用。学校积极资助博士生到世界一流大学开展交流与合作研究，超过 60%的博士生有海外访学经历。清华于 2011 年设立了发展中国家博士生项目，鼓励学生到发展中国家亲身体验和调研，在全球化背景下研究发展中国家的各类问题。

在博士学位评定方面，权力要进一步下放，学术判断应该由各领域的学者来负责。院系二级学术单位应该在评定博士论文水平上拥有更多的权力，也应担负更多的责任。清华大学从 2015 年开始把学位论文的评审职责授权给各学位评定分委员会，学位论文质量和学位评审过程主要由各学位分委员会进行把关，校学位委员会负责学位管理整体工作，负责制度建设和争议事项处理。

全面提高人才培养能力是建设世界一流大学的核心。博士生培养质量的提升是大学办学质量提升的重要标志。我们要高度重视、充分发挥博士生教育的战略性、引领性作用，面向世界、勇于进取，树立自信、保持特色，不断推动一流大学的人才培养迈向新的高度。

清华大学校长

2017 年 12 月

丛书序二

以学术型人才培养为主的博士生教育，肩负着培养具有国际竞争力的高层次学术创新人才的重任，是国家发展战略的重要组成部分，是清华大学人才培养的重中之重。

作为首批设立研究生院的高校，清华大学自 20 世纪 80 年代初开始，立足国家和社会需要，结合校内实际情况，不断推动博士生教育改革。为了提供适宜博士生成长的学术环境，我校一方面不断地营造浓厚的学术氛围，一方面大力推动培养模式创新探索。我校从多年前就已开始运行一系列博士生培养专项基金和特色项目，激励博士生潜心学术、锐意创新，拓宽博士生的国际视野，倡导跨学科研究与交流，不断提升博士生培养质量。

博士生是最具创造力的学术研究新生力量，思维活跃，求真求实。他们在导师的指导下进入本领域研究前沿，吸取本领域最新的研究成果，拓宽人类的认知边界，不断取得创新性成果。这套优秀博士学位论文丛书，不仅是我校博士生研究工作前沿成果的体现，也是我校博士生学术精神传承和光大的体现。

这套丛书的每一篇论文均来自学校新近每年评选的校级优秀博士学位论文。为了鼓励创新，激励优秀的博士生脱颖而出，同时激励导师悉心指导，我校评选校级优秀博士学位论文已有 20 多年。评选出的优秀博士学位论文代表了我校各学科最优秀的博士学位论文的水平。为了传播优秀的博士学位论文成果，更好地推动学术交流与学科建设，促进博士生未来发展和成长，清华大学研究生院与清华大学出版社合作出版这些优秀的博士学位论文。

感谢清华大学出版社，悉心地为每位作者提供专业、细致的写作和出

版指导，使这些博士论文以专著方式呈现在读者面前，促进了这些最新的优秀研究成果的快速广泛传播。相信本套丛书的出版可以为国内外各相关领域或交叉领域的在读研究生和科研人员提供有益的参考，为相关学科领域的发展和优秀科研成果的转化起到积极的推动作用。

感谢丛书作者的导师们。这些优秀的博士学位论文，从选题、研究到成文，离不开导师的精心指导。我校优秀的师生导学传统，成就了一项项优秀的研究成果，成就了一大批青年学者，也成就了清华的学术研究。感谢导师们为每篇论文精心撰写序言，帮助读者更好地理解论文。

感谢丛书的作者们。他们优秀的学术成果，连同鲜活的思想、创新的精神、严谨的学风，都为致力于学术研究的后来者树立了榜样。他们本着精益求精的精神，对论文进行了细致的修改完善，使之在具备科学性、前沿性的同时，更具系统性和可读性。

这套丛书涵盖清华众多学科，从论文的选题能够感受到作者们积极参与国家重大战略、社会发展问题、新兴产业创新等的研究热情，能够感受到作者们的国际视野和人文情怀。相信这些年轻作者们勇于承担学术创新重任的社会责任感能够感染和带动越来越多的博士生，将论文书写在祖国的大地上。

祝愿丛书的作者们、读者们和所有从事学术研究的同行们在未来的道路上坚持梦想，百折不挠！在服务国家、奉献社会和造福人类的事业中不断创新，做新时代的引领者。

相信每一位读者在阅读这一本本学术著作的时候，在吸取学术创新成果、享受学术之美的同时，能够将其中所蕴含的科学理性精神和学术奉献精神传播和发扬出去。

清华大学研究生院院长

2018 年 1 月 5 日

导师序言

时间序列是将特定研究对象以数值、向量、物体等形式，按照时间顺序记录的数据。时间序列分析通过研究已经记录的历史数据，提炼统计规律，并用以对未来序列进行预测，是统计学中的一个重要分支。随着技术的推进与发展，现在收集到的时间序列数据往往呈现维度高、相关性强、时变性强等复杂特征。复杂时间序列数据对分析和建模提出了新的挑战，亟须新的统计推断方法。本书聚焦复杂时间序列数据的统计理论和实际应用，研究了三个极具代表性的统计推断问题，即时间序列分布函数的统计推断，函数型时间序列的统计推断，以及局部平稳时间序列多步向前预测区间的统计推断。

第 2 章针对严平稳时间序列的分布函数建立了四种同时置信带。其中一个有趣的应用是通过构造的同时置信带来检验股票每日回报率分布函数的整体形状（例如重尾和尖峰态）。第 2 章在最宽泛的假设下得到了渐近结论，并检验了标准普尔 500 指数序列 (1950 年 1 月至 2018 年 8 月) 的分布函数。一个令人惊讶的发现是其分布函数可以是自由度大于 2 的多个学生分布，甚至是正态分布。本书提出的同时置信带是现有针对时间序列分布函数形状检验的唯一理论可靠的工具。该工作使用了核光滑、强混合性和随机过程收敛等结果，并于 2019 年在北大-清华统计学论坛获优秀墙报奖。

第 3 章对于平稳函数型时间序列的均值函数提出了一种渐近正确的同时置信带。具有时间相依性的函数型数据在科学研究中频繁出现，例如脑电图和心电图信号等，它们被称为“函数型时间序列”。第 3 章将轨迹间依赖性建模为取值为 L^2 空间中的无穷移动平均，记作 $\mathrm{FMA}(\infty)$。在函数型主成分得分和测量误差的基本矩条件假设下，本书建立了均值函

数 B 样条估计量的默示有效性，并基于此推导了同时置信带。对于脑电图信号这一函数型时间序列，该工作表明其均值函数实际上可以表示为稀疏的傅里叶级数，因为脑电图时间序列均值函数的三角级数估计量被包含在低置信水平的同时置信带中。这是目前唯一针对离散观测、存在测量误差的函数时间序列的同时置信带工作，其理论推导运用了 B 样条平滑、高斯过程部分和强逼近等复杂技巧。该工作于 2020 年荣获国际数理统计协会 (Institute of Mathematical Statistics，IMS) 颁发的汉南研究生旅行奖 (Hannan Graduate Student Travel Award)，作者李杰博士是当年唯一来自中国的获奖者。

第 4 章构造了局部平稳时间序列的多步向前预测区间。具体步骤是，通过 B 样条估计时变趋势函数，通过核光滑方法估计时变方差函数，在对标准化的时间序列拟合自回归模型后，得到预测残差的分位数估计，最后构造出了未来观测的预测区间。第 4 章用提出的新方法分析了西安市 2013 年 1 月至 2020 年 7 月大气污染物的浓度数据，发现本书提出的预测区间精度高于季节性 ARIMA 方法的预测区间，从而证明了所提方法的优越性。该工作于 2021 年被国际统计学会 (International Statistical Institute，ISI) 认定解决了一个对广大发展中国家具有实际意义的应用统计问题，荣获每两年颁发一次的国际统计学会简 • 丁伯根奖一等奖（ISI Jan Tinbergen Award Division • A First Prize），这也是此奖项的一等奖首次授予华人统计学者。

本书在理论分析中灵活运用了非参数统计的核估计与 B 样条估计方法，高斯强逼近、随机过程的弱收敛，以及时间序列的混合性质等多种技巧和工具。本书提出的方法适用于经济、生物、环境等诸多领域的实际数据，展现了本书成果对相关领域产业应用的重要意义，很好地体现了现代统计学研究方向的交叉性质。本书结构框架清晰，学术表达专业严谨，写作规范，希望能够给相关领域的研究带来一定的启示。

杨立坚

2023 年 8 月

摘　要

时间序列模型广泛应用于计量经济学、金融学、生物统计学、工业计量学等领域。本书主要研究了复杂时间序列的理论性质和实际应用，包括对时间序列的分布函数、函数型时间序列，以及局部平稳时间序列多步向前预测区间的统计推断。

关于时间序列的分布函数，本书从时间序列的实现中简单随机抽样，构造了基于核分布函数和经验分布函数的柯尔莫哥洛夫-斯米尔诺夫类型的同时置信带。最终构造的所有同时置信带与基于独立同分布样本构造出的同时置信带相比，有相同的柯尔莫哥洛夫-斯米尔诺夫极限分布，该结果在多个时间序列的模拟实验中得到了验证。实证中检验了标准普尔 500 指数股票从 1950 年至 2018 年日收益率的分布函数，发现自由度大于等于 3 的学生分布或经过调节的正态分布都是其可能的分布。这些发现对长期以来认为股票每日回报率数据分布是重尾和尖峰态的观点提出了挑战。

对于具有无穷滑动平均结构的平稳函数型时间序列，本书研究了其均值函数的统计推断。本书用 B 样条估计了按照时间顺序排列的轨迹，并用其构造了均值函数的两步估计量。在较为一般的假设条件下，B 样条估计量具有“默示有效”的渐近最优性，即它渐近等价于一个“不可获得”的估计量：基于无测量误差的真实轨迹估计的样本均值。这种“默示有效性”允许本书构造出渐近正确的均值函数的同时置信带。模拟结果充分证实了渐近理论。该方法为脑电图序列“可能具有三角级数形式的均值函数”这一假设提供了强有力的证据。

为了构建局部平稳时间序列多步向前的预测区间，本书将等距设计的非参数回归模型扩展到时间序列上。本书提出用 B 样条方法估计趋势

函数，用核光滑方法估计方差函数，在拟合误差的自回归模型并获得预测残差的分位数后，建立起多步向前的未来观测的预测区间。该方法在数值模拟和西安市 8 年每日空气污染物浓度的数据实例中得到了验证。最终结果表明，本书提出的方法因更高的预测精度和更广泛的适用性而优于其他方法。

关键词：同时置信带；函数型时间序列；预测区间；核光滑；B 样条

Abstract

Time series are widely used in econometrics, finance, biostatistics, industrial metrology and other fields. This book studies the theoretical property and practical application of complex time series, including statistical inference for the distribution function of time series, functional time series and multi-step-ahead prediction interval of locally stationary time series.

For the distribution function of time series, its Kolmogorov-Smirnov type simultaneous confidence bands (SCBs) are proposed based on simple random samples (SRSs) drawn from realizations of time series, together with smooth SCBs using kernel distribution estimator (KDE) instead of empirical cumulative distribution function of the SRS. All SCBs are shown to enjoy the same limiting distribution as the standard Kolmogorov – Smirnov for I.I.D. sample, which is validated in simulation experiments on various time series. Computing these SCBs for the standardized S&P 500 daily returns data leads to some rather unexpected findings, i.e., student's t-distributions with degrees of freedom no less than 3 and the normal distribution are all acceptable versions of the standardized daily returns series' distribution, with proper rescaling. These findings present challenges to the long held belief that daily financial returns distribution is fat-tailed and leptokurtic.

For stationary functional time series data with infinite moving average structure, statistical inference for its mean function is investigated. B-spline estimation is proposed for the temporally ordered trajectories

of the functional moving average (FMA), which are used to construct a two-step estimator of the mean function. Under mild conditions, the B-spline mean estimator enjoys oracle efficiency in the sense that it is asymptotically equivalent to the infeasible estimator which is the sample mean of all trajectories observed entirely without errors. This oracle efficiency allows for the construction of SCB for the mean function which is asymptotically correct. Simulation results strongly corroborate the asymptotic theory. Using the SCB to analyze an electroencephalogram time series reveals strong evidence of trigonometric form mean function.

To construct multi-step-ahead prediction interval of locally stationary time series, the nonparametric regression model with auto-regressive errors for equally spaced design is extended to the time series setup. A B-spline estimator for the trend function as well as a kernel estimator for the variance function are proposed. Prediction interval of multi-step-ahead future observation is also constructed after fitting the auto-regressive model of errors and obtaining the quantile of prediction residuals. The proposed method is illustrated by various simulation studies and an example of air pollutants data, which contains 8 years of daily air pollutants concentration in Xi'an. Final results demonstrate that the proposed method outperforms others for its higher prediction accuracy and wider applicability.

Key Words: simultaneous confidence band; functional time series; prediction interval; kernel smoothing; B-spline

目 录

第 1 章 引言 ····· 1
1.1 非参数统计方法 ····· 1
1.2 时间序列的分布函数 ····· 2
1.3 函数型时间序列 ····· 4
1.4 时间序列的预测区间 ····· 6
1.5 内容和结构 ····· 8

第 2 章 时间序列分布函数的同时置信带 ····· 10
2.1 主要结果 ····· 13
2.2 实施方法 ····· 15
2.3 数值模拟 ····· 16
2.3.1 基本数值模拟 ····· 16
2.3.2 与参数型同时置信带的比较 ····· 20
2.4 实际数据分析 ····· 24
2.5 证明 ····· 25
2.5.1 预备引理 ····· 26
2.5.2 定理 2.1 的证明 ····· 27
2.5.3 定理 2.2 所用引理及证明 ····· 28

第 3 章 函数型时间序列的统计推断 ····· 33
3.1 B 样条估计量及其渐近理论 ····· 35
3.2 分解 ····· 38
3.3 实施方法 ····· 40
3.3.1 节点数选择 ····· 40

3.3.2 协方差估计 ····· 40
3.3.3 分位数估计 ····· 41
3.4 数值模拟 ····· 41
3.5 实际数据分析 ····· 44
3.6 证明 ····· 46
3.6.1 预备引理 ····· 46
3.6.2 定理 3.1 的证明 ····· 56
3.6.3 定理 3.2 的证明 ····· 59

第 4 章 局部平稳时间序列的多步向前预测区间 ····· 61
4.1 预测区间的构造方法 ····· 62
4.1.1 估计趋势函数 $m(\cdot)$ ····· 62
4.1.2 估计方差函数 $\sigma^2(\cdot)$ ····· 63
4.1.3 自回归系数估计 ····· 63
4.1.4 建立 Y_{T+k} 的预测区间 ····· 63
4.2 实施方法 ····· 65
4.3 数值模拟 ····· 66
4.4 实证分析 ····· 73
4.4.1 探索性数据分析 ····· 73
4.4.2 基于季节性 ARIMA 模型预测空气污染物浓度 ····· 76
4.4.3 基于所提出的方法预测空气污染物浓度 ····· 79

第 5 章 工作总结与未来展望 ····· 84

参考文献 ····· 85

在学期间完成的相关学术成果 ····· 89

致谢 ····· 90

Contents

Chapter 1 Introduction 1
1.1 Nonparametric Statistical Methods 1
1.2 Distribution Function of Time Series 2
1.3 Functional Time Series 4
1.4 Prediction Intervals for Time Series 6
1.5 Content and Structure 8

Chapter 2 SCBs for Time Series Distribution Function 10
2.1 Main Results 13
2.2 Implementation 15
2.3 Simulation 16
2.3.1 General Simulation Studies 16
2.3.2 Comparison with Parametric SCB 20
2.4 Real Data Analysis 24
2.5 Proof 25
2.5.1 Preliminary Results 26
2.5.2 Proof of Theorem 2.1 27
2.5.3 Proof of Theorem 2.2 28

Chapter 3 Statistical Inference for Functional Time Series 33
3.1 B-spline Estimator and Its Asymptotic Properties 35
3.2 Decomposition 38
3.3 Implementation 40

3.3.1 Knots Selection ····· 40
3.3.2 Covariance Estimation ····· 40
3.3.3 Estimating the Percentile ····· 41
3.4 Simulation ····· 41
3.5 Real Data Analysis ····· 44
3.6 Proof ····· 46
3.6.1 Preliminary Results ····· 46
3.6.2 Proof of Theorem 3.1 ····· 56
3.6.3 Proof of Theorem 3.2 ····· 59

Chapter 4 Multi-step-ahead Prediction Interval for Locally Stationary Time Series ····· 61
4.1 Methodology ····· 62
4.1.1 Estimating the Trend Function $m(\cdot)$ ····· 62
4.1.2 Estimating the Variance Function $\sigma^2(\cdot)$ ····· 63
4.1.3 Autoregressive Coefficients Estimation ····· 63
4.1.4 Constructing PI for Y_{T+k} ····· 63
4.2 Implementation ····· 65
4.3 Simulation ····· 66
4.4 Real Data Analysis ····· 73
4.4.1 Pre-analysis Exploration ····· 73
4.4.2 Forecasts of Air Pollutants Concentration from Seasonal ARIMA Model ····· 76
4.4.3 Forecasts of Air Pollutants Concentration by the Proposed Method ····· 79

Chapter 5 Summary and Future Prospects ····· 84

References ····· 85

Relevant Academic Achievements ····· 89

Acknowledgements ····· 90

第 1 章　引　　言

时间序列分析是统计学中的一个重要研究方向，其在经济金融、生物医学、环境和工业等领域有着广泛的应用。本书主要研究了复杂时间序列的理论性质和预测方法，包括对时间序列分布函数、函数型时间序列，以及时间序列预测区间的统计推断。

本书在研究以上三个问题时均使用了非参数统计方法，下文先简要介绍一些非参数统计的基础知识，再分别介绍本书研究内容的具体背景、现有模型和现有方法的不足。

1.1　非参数统计方法

非参数回归是估计未知函数的一种重要方法。与传统的参数方法相比，它无需事先假设回归函数的形式，能更好地从实际出发，拟合复杂函数，刻画非线性关系，提高了统计模型的适应性。常用的非参数分析方法有核光滑法和样条光滑法。核光滑法主要包括 Nadaraya-Watson 估计和局部多项式估计，它们都是局部加权平均值的估计方法。而样条光滑法中的样条估计量是一种全局的估计量，通过一次计算优化即可得到。B 样条因其易于计算和概念简单，广泛用于非参数统计中，详见 De Boor[1] 和 Lorentz 等[2] 的文献。下面对 B 样条进行简单介绍。

为了描述 B 样条函数，定义 $\{t_\ell\}_{\ell=1}^{J_s}$ 为一个等距点序列，其中 $t_\ell = \ell/(J_s+1)$，$0 \leqslant \ell \leqslant J_s+1$。$0 = t_0 < t_1 < \cdots < t_{J_s} < 1 = t_{J_s+1}$ 为内部节点，它把 $[0,\ 1]$ 分成了 (J_s+1) 个相等的子区间，$I_\ell = [t_\ell,\ t_{\ell+1})$，$\ell = 0,\ \cdots,\ J_s-1$ 且 $I_{J_s} = [t_{J_s},\ 1]$。令 $\mathcal{H}^{(p-2)} = \mathcal{H}^{(p-2)}[0,\ 1]$ 为 $I_\ell,\ \ell = 0,\ \cdots,\ J_s$ 上的多项式样条空间。它包含所有在子区间 I_ℓ 上

是 $(p-1)$ 次多项式且在 $[0,\ 1]$ 上是 $(p-2)$ 次连续可微的函数。定义 $\{B_{\ell,\,p}(\cdot),\ 1 \leqslant \ell \leqslant J_s+p\}$ 为 $\mathcal{H}^{(p-2)}$ 空间的 p 阶 B 样条基函数，所以 $\mathcal{H}^{(p-2)} = \left\{ \sum\limits_{\ell=1}^{J_s+p} \lambda_{\ell,\,p} B_{\ell,\,p}(\cdot) \,\middle|\, \lambda_{\ell,\,p} \in \mathbb{R} \right\}$。通过估计系数 $\lambda_{\ell,\,p}$ 就可以在 $\mathcal{H}^{(p-2)}$ 空间上找到对未知函数的最佳逼近。

本书对函数的统计推断进行了多次研究，不同于单点随机变量使用的置信区间，这里使用的工具为同时置信带（simultaneous confidence band，SCB）。从动态的角度看，同时置信带被视作一个可移动的置信区间在未知函数定义域上滑动后的轨迹，刻画了未知函数整体的变化性，从而可以对未知函数进行全局的统计推断。同时置信带被广泛应用于统计学不同的研究领域中，其中包括非参数回归（Song 等[3]、Wang 等[4]、Wang[5]、Cai 等[6]、Zhang 等[7]）、半参数降维（Gu 等[8]、Zheng 等[9]）、函数型数据分析（Cardot 等[10]、Degras[11]、Cao 等[12]、Ma 等[13]、Cardot 等[14]、Song 等[15]、Zheng 等[16]、Gu 等[17]、Cao 等[18]、Choi 等[19]、Wang 等[20]、Yu 等[21]）、时间序列误差的分布函数的估计（Wang 等[22]、Kong 等[23]）。

1.2 时间序列的分布函数

随机变量的概率分布函数包含关于随机变量充足的信息，例如股票收益回报率的分布函数。概率分布通常由峰度、对称性、正态性和重尾性等衡量，但存在很多未被证实的说法。其中一种说法是股票收益回报率的分布函数是重尾且尖峰态的，因为它的经验分布“看起来”不像是正态的，其样本峰度远大于 3，并且经典的正态性检验会得到“拒绝”的结论。

以图 1.1 所示的 1950 年 1 月 3 日到 2018 年 8 月 28 日的标准普尔 500 指数股票每日收益回报率为例。这个长度为 17276 的时间序列显然不是平稳的，它在最近 10 年内的变化幅度更大。通过调整方差趋势将其标准化 (变成平稳序列) 并进行进一步分析，见图 1.2。图 1.3 展示了标准化后平稳序列的经验分布函数和 99%柯尔莫哥洛夫-斯米尔诺夫 (Kolmogorov-Smirnov) 同时置信带，以及均值和方差分别等于样本均值和样本方差的正

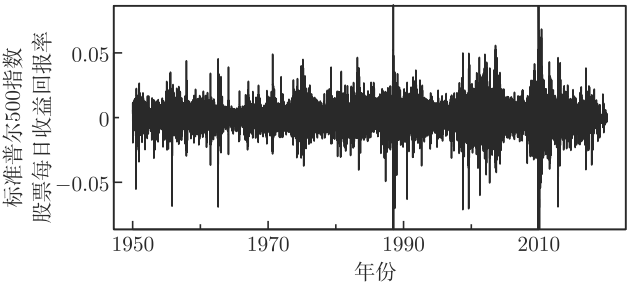

图 1.1 标准普尔 500 指数股票每日收益回报率的散点图

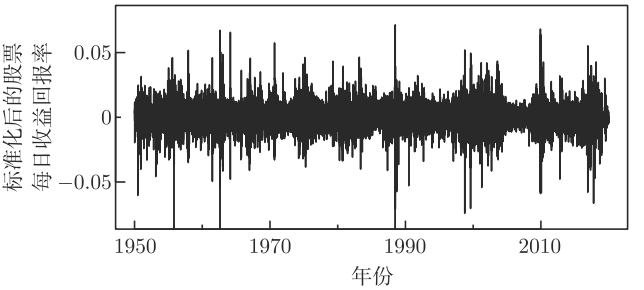

图 1.2 标准化后的股票每日收益回报率的散点图

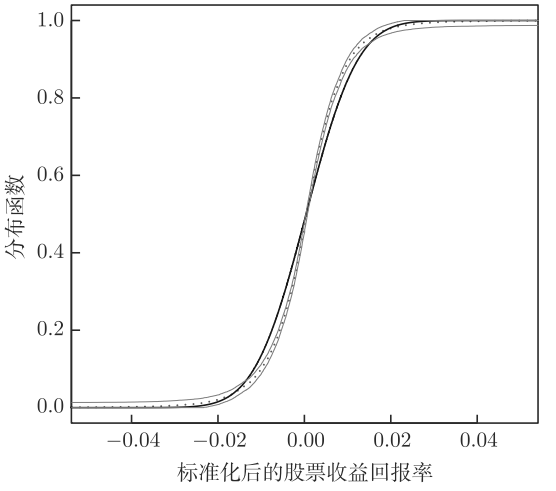

图 1.3 标准化后的股票收益回报率的经验分布函数，99%柯尔莫哥洛夫-斯米尔诺夫同时置信带，以及正态分布函数（前附彩图）

蓝色虚线是经验分布函数，红色实线是 99%的柯尔莫哥洛夫-斯米尔诺夫同时置信带，黑色实线是均值和方差等于样本均值和样本方差的正态分布函数。

态分布函数。该图很直观地说明了“看起来不正态”的情况，因为正态分布函数落在了同时置信带的外面。但经典的正态性检验方法，例如基于柯尔莫哥洛夫-斯米尔诺夫同时置信带的检验，仅适用于独立同分布的观测，而独立性假设对时间序列是难以满足的，因此需要更直接、更可靠的方法来检验时间序列分布函数的形状，验证各种假设的有效性。

1.3 函数型时间序列

近年来，函数型数据分析已成为统计学热门的研究领域。函数型数据也被称作“曲线数据”，广泛出现在生物医学、流行病学和社会科学等研究中，由对每个个体在一段时期内进行多次观测收集得到。它把多元数据统计分析扩展到更复杂、信息量更大的曲线数据分析（Ferraty 等[24]、Silverman 等[25]、Ramsay 等[26]、Hsing 等[27] 和 Kokoszka 等[28]）。从数学领域上讲，经典的函数型数据由 n 条曲线 $\{\eta_t(\cdot)\}_{t=1}^n$ 组成，对应 n 个个体。其中，第 t 个个体的曲线 $\eta_t(\cdot)$ 是一个连续的随机过程，并且和一个标准的随机过程 $\eta(\cdot)$ 同分布。这些曲线 $\{\eta_t(\cdot)\}_{t=1}^n$ 扮演着与经典统计中单变量或多变量的随机观测相同的角色，因此可以根据这些随机曲线预测其他数值型或名义型的响应变量，或者在更基本的层面研究这些曲线的位置和尺度参数。后者包含 $\eta(\cdot)$ 的均值函数 $m(\cdot)=\mathbb{E}\{\eta(\cdot)\}$，方差函数 $G(x，x')=\mathrm{Cov}\{\eta(x)，\eta(x')\}$。Cao 等[12,18] 和 Zheng 等[16] 基于不同的极限分布研究了 $m(\cdot)$ 和 $G(\cdot，\cdot)$ 逐点的正态置信区间和同时置信带。

函数型数据分析的很多工作都是基于“原始的”函数型数据 $\{Y_{tj}\}$ 开展的。其中，Y_{tj} 代表第 t 条曲线 $\eta_t(\cdot)$ 在第 j 个位置 X_{tj} 上的观测值，并且带有观测误差 $\sigma(X_{tj})\varepsilon_{tj}$，即

$$Y_{tj}=\eta_t(X_{tj})+\sigma(X_{tj})\varepsilon_{tj}，\ 1\leqslant t\leqslant n，\ 1\leqslant j\leqslant N_t \tag{1.1}$$

所以离散的原始数据不是“光滑数据”$\{\eta_t(\cdot)\}_{t=1}^n$ 的集合。在稠密观测的情况下，曲线 $\{\eta_t(\cdot)\}_{t=1}^n$ 能够被一一估计，产生和 $\{\eta_t(\cdot)\}_{t=1}^n$ 很像的估计量。具体来说，“原始数据”有以下形式：

$$Y_{tj}=\eta_t\left(\frac{j}{N}\right)+\sigma\left(\frac{j}{N}\right)\varepsilon_{tj}，\ 1\leqslant t\leqslant n，\ 1\leqslant j\leqslant N \tag{1.2}$$

其中，N 和 n 都趋向于无穷。Cao 等[12] 得到了样条估计量 $\{\widehat{\eta}_t(\cdot)\}_{t=1}^n$，被称作“伪光滑数据”，它在数据分析中可以代替 $\{\eta_t(\cdot)\}_{t=1}^n$。不失一般性，假设函数 $\eta(\cdot)$ 和 $\{\eta_t(\cdot)\}_{t=1}^n$ 都定义在 $[0,\ 1]$ 上，$G(\cdot,\ \cdot)$ 定义在 $[0,\ 1]^2$ 上。曲线 $\{\eta_t(\cdot)\}_{t=1}^n$ 可被分解为 $\eta_t(x)=m(x)+\xi_t(x)$，其中，$m(\cdot)$ 在 $[0,\ 1]$ 上连续，$\xi_t(x)$ 是 x 在第 t 条曲线上的一个小范围变化，可视作一个样本路径连续的随机过程，且 $\mathbb{E}\xi_t(x)=0$，$\mathbb{E}\max_{x\in[0,\ 1]}\xi_t^2(x)<\infty$，协方差函数 $G(x,\ x')=\mathrm{Cov}\{\xi_t(x),\ \xi_t(x')\}$。

根据 Hsing 等[27] 的结论，存在 $G(\cdot,\ \cdot)$ 的特征值 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant 0$，满足 $\sum\limits_{k=1}^{\infty}\lambda_k<\infty$，相应的特征函数 $\{\psi_k\}_{k=1}^{\infty}$ 组成了 $L^2[0,\ 1]$ 空间的一组正交基，使得 $G(x,\ x')=\sum\limits_{k=1}^{\infty}\lambda_k\psi_k(x)\psi_k(x')$，$\int G(x,\ x')\ \psi_k(x')\,\mathrm{d}x'=\lambda_k\psi_k(x)$。随机过程 $\eta(x)$，$x\in[0,\ 1]$，具有著名的 Karhunen-Loève L^2 展开形式：$\eta(x)=m(x)+\sum\limits_{k=1}^{\infty}\xi_k\phi_k(x)$，其中随机系数 $\{\xi_k\}_{k=1}^{\infty}$ 称作“函数型主成分”（functional principle components，FPC）得分，它们是均值为 0、方差为 1 的随机变量。调节过的特征函数 ϕ_k 即被称作“函数型主成分”，且对于 $k\geqslant 1$，$\phi_k=\sqrt{\lambda_k}\psi_k$。

估计均值函数通常是函数型数据分析的第一步，Ma 等[13] 和 Zheng 等[16] 研究了稀疏型函数型数据均值函数的渐近理论和应用，Cao 等[12] 研究了稠密型函数型数据均值函数的同时置信带。Cao 等[12] 和 Cao 等[18] 的研究中存在的一个明显缺点是假设每条曲线上的观测数量被曲线数量所控制，即对于某些 $\theta>0$，$N=\mathcal{O}\left(n^{\theta}\right)$。这种约束实际上是不合理的，它限制了每条曲线有更密集的观测，而研究者总希望有更大的观测数量 N。因为无论曲线数量 n 是大还是小，更大的 N 都意味着测量精度的提高（考虑极限情况 $N=\infty$，即观测点在整个范围内任意密集）。所以将假设改为更合乎逻辑的 $n=\mathcal{O}\left(N^{\theta}\right)$，也即 n 的增长速度取决于由 N 设定的精度水平。

目前，很多已有工作都限制 $\{\eta_t(\cdot)\}_{t=1}^n$ 是随机过程 $\eta(\cdot)$ 的独立同分布的样本，这显然不符合函数型时间序列的情况。一个有趣的例子是闭眼静息态下连续测量的脑电图信号数据。被试者要接受 5min 的测试，从头部 32 个不同位置上以 1000Hz 的频率收集脑电图信号。把第 6 个位置上的脑电图信号连续分成 400 段，每段包含间隔为 0.001s 的 $N=500$ 个

脑电图信号。图 1.4 展示了从 400 条光滑曲线 $\{\widehat{\eta}_t(\cdot)\}_{t=1}^{400}$ 中随机抽取的 5 个“伪光滑数据”。

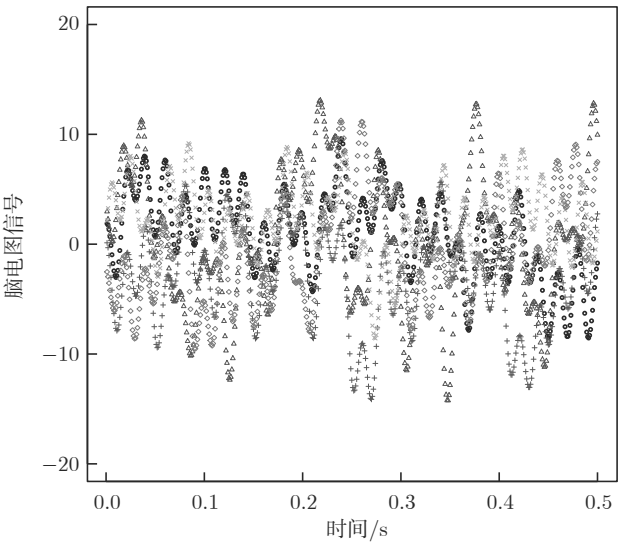

图 1.4　脑电图信号时间序列中的 5 条光滑曲线（前附彩图）

Horvath 等[29] 提出了检验存在时间相关性的函数型数据的两样本均值函数是否相等的方法，但前提是所有 $\{\eta_t(\cdot)\}_{t=1}^{n}$ 能被完整观测。Chen 等[30] 建立了函数型时间序列均值函数的同时置信带，但其理论证明存在两大问题。首先，它要求正特征值的数量是有限的，限制了其适用范围。其次，它没有对函数型主成分得分做出明确的假设，特别是其要求的自然独立条件不能保证所有函数型主成分得分 ξ_k，$k=1$，2，$\cdots$ 的独立性，但所有函数型主成分得分的独立性是保证 ξ_{tk} 部分和高斯强逼近结论成立的重要条件。只有所有的 ξ_{tk}，$k=1$，2，$\cdots$，k_n 是联合独立的，才能保证它们的线性组合也是高斯分布的。因此，需要更可靠、更合适的方法来解决函数型时间序列均值函数的估计问题。

1.4　时间序列的预测区间

预测是时间序列分析中的一项重要内容，其在环境学、经济学和其他学科有广泛的应用。预测区间（prediction interval）指对未来观测以一

定概率落入的区间的估计，是实现统计推断的一种有效工具（Brockwell 等[31] 和 Fan 等[32]）。

最近许多论文研究了不同时间序列模型的预测区间。特别地，Thombs 等[33] 将非参数自助法运用于自回归的预测中。Wang 等[22] 和 Kong 等[23] 基于使用 Yule－Walker 方法估计自回归系数后得到的残差，分别提出了自回归时间序列中关于未观测到误差的分布函数的核估计量，以及多步向前预测误差的分布函数的核估计量。上面三篇文章都成功地建立了多步向前的预测区间，但只是基于简单的 AR(p) 模型，远不够拟合大部分实际数据。Aneiros-Pérez 等[34] 从非参数角度，即基于同方差和自助法残差，使用函数型数据的方法处理时间序列预测。De Livera 等[35] 将 Box–Cox 变换、具有时变系数的傅里叶表示运用到复杂的季节性时间序列中，并在正态误差的假设下得到了单点预测和预测区间的解析表达式。尽管上述模型的统计推断理论已经发展成熟，但预测效果在不同场景中表现各异。有时这些模型不足以解释某些数据的复杂结构，存在由模型误设而引起的估计量非光滑的风险。此外，这些模型通常严格假设短期预测中的残差分布是正态或渐近正态的，导致在实际数据分析中可能会出现比正常情况更宽的预测区间，详见 4.4 节。

为了解决上述问题，更实际的做法是假设一个缓慢变化的随机结构，即局部平稳模型，见 Dahlhaus[36] 的文献。Dette 等[37] 提出了有光滑变化趋势的局部平稳过程的高维协方差矩阵的估计量，并用该统计量得到了非平稳时间序列相合的预测量。该文提出的预测量不依赖于拟合一个自回归模型或者衰减的趋势，而是更多地关注单点预测，不是预测区间。目前，关于局部平稳时间序列预测区间的文献很少，唯一的一篇出自 Das 等[38]。该文章在模型无关和模型依赖两种情况下，建立了局部平稳时间序列的一步向前预测和置信区间。他们使用自助法构造预测区间，从而缺乏对误差分布函数的估计，未能建立多步向前的预测区间。

值得注意的是，一些常用的模型检验方法（如留一法和交叉验证法）在时间序列的框架下会失效，因为时间序列中的变量存在时序相关性，测试集和训练集不能随意拆分。而衡量预测区间的表现是一种很好的替代方法。正如 Kong 等[23] 所指出，一个“理想”的置信区间需满足以下要求：首先它应该是准确的，即预测区间中包含未知量的概率应该接近于设

定的置信水平 $1-\alpha$；其次，它应具备有效性，即区间足够窄，能有效确定未知量的范围。因此，可以通过比较预测区间对测试集中真实值的覆盖率或者探究预测区间的长度和边界来评价预测区间的表现。

1.5 内容和结构

本书主要研究了时间序列分布函数的同时置信带的构造、平稳函数型时间序列均值函数的统计推断，以及局部平稳时间序列多步向前预测区间的构造。具体内容如下：

第 2 章提出了基于从时间序列中简单随机抽样，构造其分布函数的核分布函数和经验分布函数的柯尔莫哥洛夫-斯米尔诺夫类型的同时置信带；证明了该同时置信带和基于独立同分布样本的同时置信带极限分布相同，从而可以用来检验关于时间序列分布函数的各种假设。该方法理论可靠，易于实施，具有广泛的应用价值。在数值模拟中，研究了不同置信带在不同时间序列情形下的表现，结果表明本书提出的置信带都具有良好的渐近性质。最后，给出了基于标准普尔 500 指数股票每日回报率时间序列数据的实证案例，研究表明在显著性水平为 0.05 的情况下，其分布函数可能是正态的，也可能是自由度大于或等于 3 的学生分布。

第 3 章研究了函数型时间序列的统计推断问题。曲线之间的时间相依性被建模为取值在 L^2 空间上的序列的移动平均。在关于函数型主成分得分和测量误差的基本矩条件假设下，证明了均值函数 B 样条估计量的默示有效性，从而推导出同时置信带。数值模拟有力地证明了渐近理论，并用该方法分析了一个脑电图信号序列，发现均值函数的三角级数估计量可以被低置信水平的同时置信带完全包含，表明脑电图信号序列的均值函数实际上用傅里叶级数表示。

第 4 章基于局部平稳时间序列的框架，研究了多步向前观测的预测方法和相应预测区间的构造。先用样条回归估计趋势函数和核回归估计方差函数对所得的近似平稳序列拟合自回归模型，再用核分布方法估计其误差的分位数后，得到了带趋势项的自回归时间序列的数据驱动多步向前预测区间。相比于季节性差分整合移动平均自回归等传统方法产生

的预测区间，本书提出的方法得到的预测区间不仅长度更窄，还具有更好的预测精度和覆盖率。最后，分析了西安市 8 年每日空气污染物浓度的数据，结果表明本书方法因更高的预测精度有更广泛的适用性。

第 5 章总结了全书的工作，并展望了未来可能的研究方向。